Centre for Computational Finance and Economic Agents

WP027-08

Working Paper Series

Mark Adams, Marco Galbiati and Simone Giansante

Emergence of tiering in large value payment systems

September 2008

University of Essex

CCFEA

www.ccfea.net

Emergence of tiering in large value payment systems

Mark Adams†, Marco Galbiati† and Simone Giansante‡

2 September 2008

Abstract

This paper develops and simulates a model of emergence of networks in an RTGS payment system.

A number of banks, faced with random streams of payment orders, choose whether to link directly to the payment system, or to use a correspondent bank. Settling payments directly on the system imposes liquidity costs, which depend on the maximum liquidity overdraft incurred during the day. On the other hand, using a correspondent entails paying a flat fee, charged by the correspondent to recoup liquidity costs and to extract a profit. We specify a protocol whereby banks sequentially choose whether to link directly to the system, or to become clients of other banks, thus generating a client-correspondent network.

We simulate this protocol, observing the emergence of different network structures. The liquidity pricing regime chosen by a Central Bank is found to affect the tiering process, determining stable network structures. A calibration exercise on data from the UK CHAPS system suggests that the model is able to generate realistic predictions, i.e. networks similar to the one observed in reality.

† Contact authors. E-mail: Marco.Galbiati@bankofengland.co.uk. The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England.

‡ CCFEA, University of Essex. E-mail: sgians@essex.ac.uk.

The authors thank various colleagues at the Bank of England for useful comment, and participants to the CCFFEA Summerschool 2007 (University of Essex) for feedback on the paper. The usual disclaimer applies.

1 Introduction

Huge amounts of money flow through large value payment systems (LVPSs). In 2006, interbank payments in the UK's CHAPS system averaged £200 billion ($400 billion) a day; the corresponding transactions in the US Fedwire system amount to about twice as much, while in the Euro area's TARGET system volumes are roughly three times as large. Considering these staggering amounts, it natural that central banks and policy makers are interested in the smooth functioning of LVPSs, devoting substantial resources to their study, design and oversight.

These large aggregate flows are only part of the picture, as the structure of LVPSs differ drastically from country to country. In the UK for example, the main system has only 14 direct, 'first-tier' members, who settle payments on behalf of about 420 other institutions. At the other extreme, the US Fedwire system has a much less tiered structure: over 9500 banks, some of which are very small, link to the system directly and settle payments on their own behalf. A number of recent studies have charted the topology of payments over these networks in detail: Soramäki et al. (2007) look at the US Fedwire system; Becher et al. (2007) consider the UK CHAPS, Lublóy (2006) study the Hungarian VIBER, while Inaoka et al. (2004) look at the Japanese payment system BOJ-NET.

What lies behind these differences? Why do certain banks join a LVPS, while others who are eligible to join make their payments via a first-tier correspondent? These questions are important for policy makers because, first, the network structure of a payment system may affect the stability and efficiency of the system itself. Second, tiering implies that a share of interbank payments does not cross the official LVPS at all, settling instead across the books of the first tier banks.[1] Here, however, we do not attempt to clarify which structure is most desirable from a central bank's perspective. This paper concentrates instead on the following questions: what determines the structure of a payment system? Can a central bank induce the formation of a particular network structure?

To answer these questions, one must consider the incentives to join the first-tier of a LVPS, versus those to remain in the second-tier. Direct membership is expensive: first, it imposes fixed costs such as fees and back office expenses to connect to the system. Second, and perhaps more importantly, a first tier bank must have sufficient liquidity to support its payment activity on a continuous basis. Indeed, most LVPSs nowadays work in RTGS (Real Time Gross Settlement) mode: if bank i owes £2 to bank j, and j owes £1 to i, both i and j must transfer the *full* amounts of their payments to their counterparty, while no netting is allowed (with netting instead, all is due is a 1£ payment from i to j).[2] The gross modality imposes high liquidity demands on the banks,

[1] This share can be large: it is estimated to be about 30% for the UK, or about £100 billion daily. For a discussion of risks involved in tiered payment systems, see Harrison et al. (2007).

[2] Untill two decades ago, most LVPSs worked on an end-of-the-day-net basis. Gross systems were introduced worldwide to eliminate the credit exposures that would otherwise build throughout the day.

exposing them to the risk of large (albeit temporaneous) funds outflows. And management of these flows represents one of the challenges for a first-tier bank.

Jackson et al. (2006) look at a bank's decision whether to become direct member of a system, or to use a correspondent. Their findings suggest the existence of economies of scale in correspondent banking, generated by two effects: *internalization* of payments and *liquidity pooling*. Internalization refers to the fact that, when a bank acts as a correspondent, payments between its clients can be settled on the bank's own books ('on us'), at a zero liquidity cost. Liquidity pooling instead is a dynamic effect: by pooling uncorrelated payment requests from different clients, the liquidity need of a correspondent bank stabilizes, implying in turn that the costs of liquidity management are lowered. Internalization and liquidity pooling in CHAPS are estimated by Lasaosa et al. (2007) in a study which relates the degree of tiering to the liquidity needs of the system.

This paper looks at how the 'internalization' and the 'pooling' affect the client - correspondent structure of a LVPS. To do so, we set up a model where a number of banks face a random stream of payment requests. Banks can execute payments on their own, by borrowing liquidity from the central bank at a cost. Alternatively, they may become customers of other banks (correspondents), which can execute payments on their behalf, thus relieving them of liquidity costs. However, correspondents charge their clients a fee to recoup costs and, possibly, to make a profit. Who becomes a correspondent, and who instead remains in the 'second tier' attaching to one correspondent or to another, is endogenously determined by a dynamic process. Making some assumptions on how corresponding services are priced, offered and accepted, we look at how the client - corespondent network evolves in time, converging to a stable state.

Our model is highly stylized. First, we assume that the timing of payments is outside the banks' control: payments are made as soon as payment *orders* are received and, in turn, orders are generated by a random process whose intensity (but *not* the precise timing) depends on bank's choices[3]. In other words, we do not consider active liquidity management on the part of banks, an issue worth in itself a stand-alone paper - see e.g. Angelini (1988), Bech *et al.* (2003). As a second important simplification, we ignore all credit risk issues that may emerge between correspondents and their clients. Our paper is therefore different from Chapman *et al.* (2008), where instead credit risk is the main driver of tiering, because then correspondent banks may assume a monitoring role. The interaction of credit risk and tiering are also studied by Harrison *et al.* (2005), by Kahn *et al.* (2005), and by Lai *et al.* (2006). We leave this important issue aside, to focus on the relationship between i) the geography of the underlying payments to be made, ii) the liquidity pricing regime chosen by the central bank, and iii) the ensuing network structure of the system.

In a nutshell, our findings are that, if the cost of liquidity is proportional to the amount borrowed, economies of scale in liquidity costs bring about concen-

[3]More precisely: a bank's payments orders are generated by a Poisson process, whose parameter depends on the 'position' of the bank in the payment network.

tration in the correspondent business, generating tiering. If instead convexities are present (as in the case of liquidity lent freely, but against (freely available) collateral), tiering is reduced. In any case, the structure of the tiered network appears heavily influenced by the pattern of the underlying payments, or the 'geography' of the payment industry.

2 Model

The model has a population of N banks, sending payments to each other over a series of days. Banks can either be direct participants in the payment system, or they can hire a correspondent bank to execute payments on their behalf. If a bank participates directly, it needs to obtain liquidity from the central bank. This has a cost, which can be interpreted as either direct central bank charging for intraday overdrafts, or the opportunity cost of posting collateral at the central bank. Instead, banks that hire a correspondent only have to pay a price for the payment service. We look at how correspondent agreements evolve in time, leading to an equilibrium network structure of the payment system.

2.1 Intraday payments

Banks are indexed by $i = 1, 2, .., N$. The model has a (potentially infinite) number of 'days', each of which is in continuous time, $t \in [0, 1]$. On any day, banks send to each other £1-payments in a random fashion: during the day, the probability of bank i making k payments to bank j in time $[t, t+\delta t]$ is $\frac{e^{-p_{ij}\delta t}(p_{ij}\delta t)^k}{k!}$; that is, payments from i to j follow a homogeneous Poisson process with parameter p_{ij}. The matrix $P = [p_{ij}]$ describes the underlying economic structure in the model.

Since a sum of Poisson processes is also a Poisson process, bank i's total outgoing (incoming) payments follow a Poisson process with rate parameter $\sum_j p_{ij}$ (with parameter $\sum_j p_{ji}$). We assume that, on average, banks do not make or receive net payments, though they may do so on any individual day. Thus

$$\forall i, \quad \sum_j p_{ij} = \sum_j p_{ji} \equiv \frac{1}{2}\lambda_i \tag{1}$$

This parameter λ_i will determine a bank's expected costs. The liquidity need of bank i at time t, denoted by $L_i(t)$, is the sum of payments sent *minus* payments received up to t.

2.2 Liquidity costs and the 'pooling effect'

Each direct participant in the payment system needs to acquire enough liquidity from the central bank to cover its liquidity needs. The costs of covering this maximum liquidity overdraft are given by a function f, so a bank i's liquidity costs each day are $f(\max_{t=0..T} L_i(t))$.

Depending on the (random) order according to which payments are made and received, $\max_{t=0..T} L_i(t)$ varies from day to day. Banks are supposed to be risk-neutral, so they make decision according to expected costs. In Appendix I, the expectation of $f(\max_{t..T} L_i(t))$ is shown to be following, increasing function of λ_i:

$$C(\lambda_i) = E(C) = \sum_{n=0}^{\infty} \frac{\lambda_i^n}{2^n n!} \sum_{m=0}^{n} f_i(m) \binom{n}{\lceil \frac{n+m}{2} \rceil} \tag{2}$$

(where $\lceil x \rceil$ is the smallest integer larger than x). These are the *expected liquidity costs*, determining a bank's choices. Fixed f, they are uniquely determined by λ_i, which is in turn determined by P.[4] The following examples show how C depends on the specification of the function f.

Example 1 *Type-I costs:*

$$f(x) = cx, \quad c > 0 \tag{3}$$

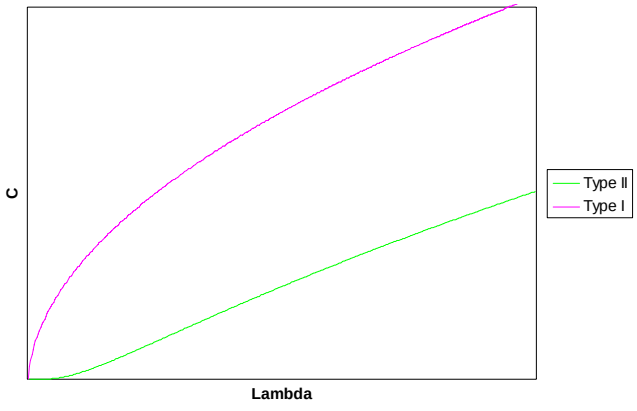
Computation shows that the resulting expected costs $C(\lambda)$ are increasing, concave, asymptotically linear, with $C'(0) < 1$ - see Figure 1.

Example 2 *Type-II costs:*

$$f(x) = \begin{cases} 0 & \text{for } x < K \\ cx, \quad c > 0 & \text{for } x \geq K \end{cases} \tag{4}$$

Computation shows that the resulting $C(\lambda_S)$ is 'S-shaped' (first flat, then as in the above case) - see Figure 1.

Figure 1 - Two cost Types



[4] By Eq. 1), pay-ins and pay-outs equally determine λ_i. So, they affect expected cost (Eq. 2) in the same way.

We use these cost-types in the simulations later on (where will clarify the reason to adopt these precise functional forms). Under both specifications, *the concavity*[5] *of the function C is the so-called 'pooling effect'*: there are economies of scale in the payment activity.

2.3 Correspondent banks and 'internalization effect'

Instead of participating directly in the payment system, a bank j may outsource its payments activity some other bank i. When this happens, bank i acts as the correspondent of bank j (which becomes client of bank i), and the following terms are agreed upon: i supports all liquidity costs deriving from j's payments, and in exchange j pays i a flat fee. A surplus is created by a correspondent agreement: first, some payments may be internalized. Second, there are economies of scale from pooling payments, as shown by Fig. 1.

In more detail, suppose bank i is correspondent for a group of banks $S = \{i, j, ...\}$. In this case, bank i's liquidity costs are determined by the payments between S and the banks outside S. Instead, payments within S can be settled by changing entries of a book and require no liquidity. That is, bank is' cost will be equal to $C(\lambda_S)$, where

$$\lambda_S = \sum_{j \notin S} \sum_{i \in S} p_{ij} + \sum_{i \notin S} \sum_{j \in S} p_{ij} \tag{5}$$

Note two that λ is sub-additive: given two groups A and B, $\lambda_{A \cup B} \leq \lambda_A + \lambda_B$. Indeed, if bank A has no payments from/to B, no payments can be internalized, so $\lambda_{A \cup B} = \lambda_A + \lambda_B$. But, if all payments made and received by B are towards A, then $\lambda_{A \cup B} = \lambda_A - \lambda_B$. In intermediate cases, $\lambda_A < \lambda_{A \cup B} < \lambda_A + \lambda_B$. *The subadditivity of λ is the so-called 'internalization effect'.*

Summing up: C is increasing but, due to internalization, adding a bank to a group S can either increase or decrease the costs of S's correspondent. As noted above (Figure 1) C is convex in a certain range so, even if $\lambda_{S \cup k} > \lambda_S$, it may still be $C(\lambda_{S \cup k}) < C(\lambda_S) + C(\lambda_k)$ i.e. a surplus may be realized by adding k to S.

2.4 Network formation

A 'network' is a partition of the N banks into groups, each with one correspondent. How do these groups form, i.e. how does the network evolve? We imagine that correspondent relationships are formed day after day, with banks accepting offers made according to the following protocol (index $t = 0, 1, 2...$ now refers to days).

1 At $t = 0$, all banks are self-settling;

2 at each $t > 0$, one randomly selected bank (say i) receives an offer from each other bank k; this is the fee k would charge i to become its correspondent;

[5] For Type-II, only above a certain lambda.

3 i chooses the best (lowest) offer, becoming client of the best offerer;

4 when a bank i becomes a client of another bank, all its clients (if any) go back to self-settling. i must pay a penalty to its previous clients for breaching their contracts.

To clarify, the selected bank i receives offers from all correspondents and all clients (a client k makes an offer considering to leave its correspondent, to become itself a correspondent for the new group $\{i,k\}$). Of course i may also maintain its role; it will do so when the expected costs of doing so are lower than any other offer.[6]

The *offer*, or fee charged by the correspondent, is determined according to the Nash Bargaining Rule (NBR). In general terms, the NBR prescribes that, if parties a and b obtain a total profit ω by signing an agreement, they divide it in two shares x_a and x_b as follows:

$$\begin{array}{l} x_a = \frac{1}{2}\left(\omega + O_a - O_b\right) \\ x_b = \frac{1}{2}\left(\omega - O_a + O_b\right) \end{array} \qquad \text{(NBR)}$$

where O_i is what i receives if the agreement is not signed. In our story, party b (the client) pays a fee, so the offer is $-x_b$; the correspondent instead takes 'the remainder'. It should be noted that these offers are 'myopic': banks do not consider that their partners might sign other contracts in the future.

Example 3 *For simplicity, for a group A we write $C(A)$ instead of $C(\lambda_A)$. Suppose that k, a self settler with no clients, makes an offer to another similar self-settler i. If the offer is rejected, the parties' profit remain $-C(k)$ and $-C(i)$. If instead the offer is accepted, the total profit for both parties is $-C(\{i,k\})$. The NBR attributes to i a profit $\frac{1}{2}[-C(\{i,k\}) + C(k) - C(i)]$, i.e. i* is asked to pay $q_{ki} = \frac{1}{2}[C(\{i,k\}) - C(k) + C(i)]$. *This is the offer that k makes to i.*

Consider now the general case: i receives an offer from k. There can be two sub-cases: i is client of some w, or it is correspondent for group S.[7] In the first sub-case, i's outside option is $O_i = -q_{wi}$. In the second sub-case, if i keeps its outside option, it bears a cost $C(S)$ but receives fees totalling $\sum_{r\in S\setminus i} q_{ir}$. So,

$$O_i = \begin{cases} -q_{wi} & \text{when } i \text{ is client of } w \\ O_i = -C(S) + \sum_{r\in S\setminus i} q_{ir} & \text{when } i \text{ is correspondent for } S \end{cases} \qquad (6)$$

To determine the joint profits to i and k, recall that, if a correspondent i leaves its group, each of its clients goes back to self-settling. Hence, each 'abandoned' bank r suffers a loss of $C(r) - q_{ir}$. Bank i is liable for this, so its defection brings about a penalty of

$$X_i = \sum_{r\in S\setminus i} \left[C(r) - q_{ir}\right]$$

[6] It's simple to see that no bank ever finds it convenient to go back to self-settling.
[7] When i is a self-settlier, $S = \{i\}$.

We suppose that i and its new correspondent k share this penalty. So when k, correspondent for group P, makes an offer to i, correspondent for group S, the profits for the new group are

$$\omega = -C\left(P \cup i\right) + \sum_{r \in P \setminus k} q_{ir} - X_i \tag{7}$$

Hence, the NBR prescribes that k charges i a fee equal to:

$$q_{ki} = \frac{1}{2}[-\omega + O_k - O_i] \ , \tag{8}$$

with ω defined in 7) and O_i defined in 6).

2.5 Dynamic properties

The abstract structure of the model is that of a 'coalitional game': we have a set of players N and, for each subset $S \subseteq N$, a payoff $C\left(\lambda_S\right)$ is given (Eqns. 5 and 2). To this coalition-form game, we attach a particular protocol (Section 2.4), specifying how coalitions form and dissolve. We don't pursue an abstract analysis of this game. However, the following fact provides theoretical ground for the simulations performed later on.

Lemma 1 *The network reaches a stable state in a finite number of steps.*

Proof. in Appendix II ■

The above Lemma ensures that no cycles are generated in our protocol, so an equilibrium *is* reached. *What* equilibrium then? The hub-and-spokes network (one bank acting as correspondent for all others), is trivially an equilibrium network.[8] However, it is easy to construct matrices P with two equilibria, both accessible from the same initial condition.[9] Because banks make decisions in a random order, one cannot speak of *the* equilibrium in general.

An analytical study of the statistic properties of these equilibria is beyond the scope of this work. Instead, we run the protocol many times using different 'seeds', for each set of non-random inputs (a matrix P and a function f). The next section illustrates the results.

3 Results

In the simulations we fix the matrix of payments P (see Sect. 2.1) and the cost function f (see Sect. 2.2). Then, we 'run' the protocol to produce a sequence of networks, starting from a situation where all banks are self-settling,up to equilibrium, i.e. until banks stop changing correspondents. Section 3.1 present some abstract examples; Section 3.2 instead calibrates the model using data from the UK CHAPS payment system.

[8] A lone correspondent internalizes all payments, and so it incurs zero costs; as a consequence, its fees cannot be undercut.

[9] If two groups have many within-group payments, and few cross-group payments, they are somewhat "far" from each another, and they can co-exist in equilibrium.

3.1 Liquidity costs, underlying payments and tiering

This section presents some abstract examples, to illustrate the relationship between i) liquidity costs, ii) the 'geography' of payments given by matrix P, and iii) the resulting network structure.

We consider the two types of liquidity cost described in Examples 1 and 2, combining them with three payment matrices described in Figure 2:

Figure 2

P': disconnected components

1	1	1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0
0	0	0	1	1	1	1	0	0	0
0	0	0	1	1	1	1	0	0	0
0	0	0	1	1	1	1	0	0	0
0	0	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	1	1

P'': complete symmetric network

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

P''': complete, asymmetric network

5	1	3	1	0	0	0	0	0	0
4	6	0	0	0	0	0	0	0	0
1	2	5	2	1	0	0	0	0	0
0	1	2	5	2	1	0	0	0	0
0	0	1	2	5	1	1	0	0	0
0	0	0	1	2	5	1	1	0	0
0	0	0	0	0	2	5	2	1	0
0	0	0	0	0	1	1	5	3	0
0	0	0	0	0	0	1	1	5	3
0	0	0	0	0	0	1	1	1	5

P' represents a disconnected payment network with three distinct payment areas. Matrix P'' instead represents a complete, symmetric payment network. Finally, in P''' banks have preferential payment partners, but the payment network is completely connected.

Matrix P' gives rise to the (rather predictable) outcome of Figure 3: three different correspondents emerges for each of the three disconnected network

components. Given symmetry, which banks becomes correspondent is randomly determined.

Figure 3: Payments P', Type-I and -II costs

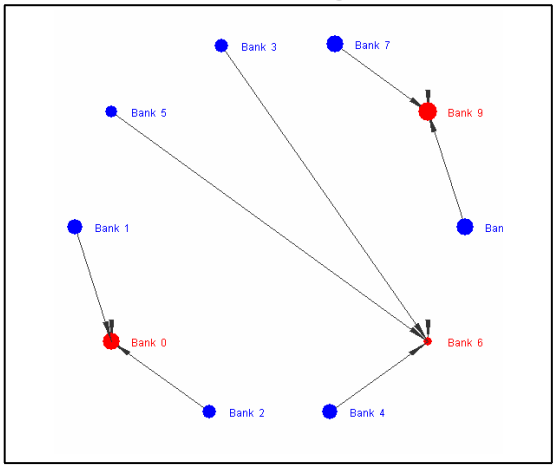
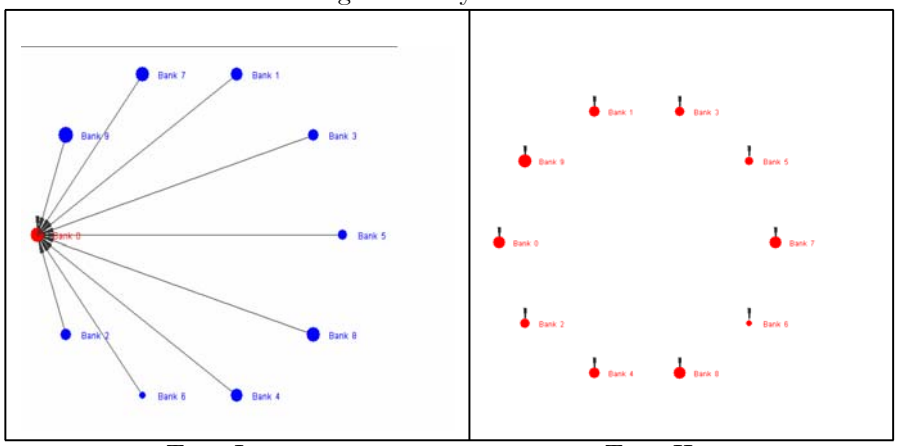


Figure 4 shows the results for matrix P''. Such an extreme case is useful to exemplify the role of cost function on the network dynamics.

Figure 4: Payments P''



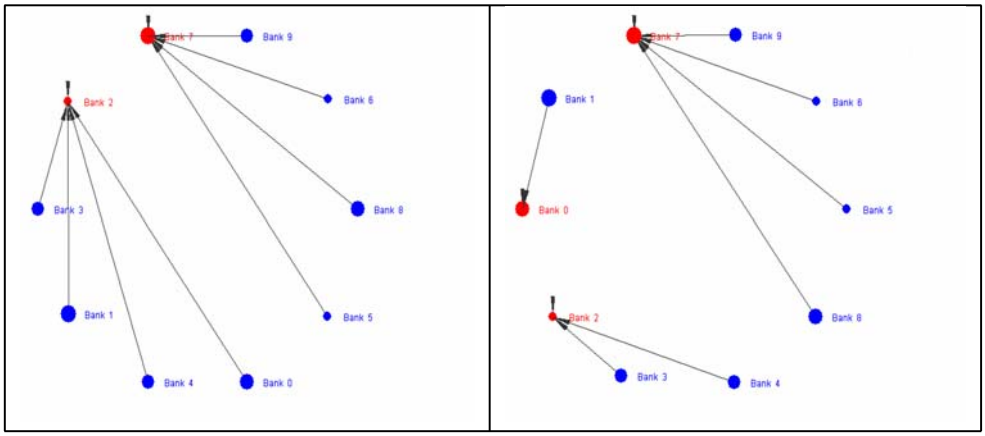
Type-I costs Type-II costs

The results in Figure 4 have the following explanation. If all banks are identical in all respects, it is intuitive that either of two cases are possible: i) one bank emerges as unique correspondent, or ii) all banks remain self-settling. The outcome is determined by the shape of the cost function: Type-I costs induce maximum tiering (unique correspondent), while the network totally unravels under Type-II costs, provided the threshold K (eq.4) falls in a certain interval. Indeed: suppose K is higher than the payments of any bank, but smaller than the payments that two banks would have to make to *others* (i.e. $\lambda_i < K < \lambda_{i\cup j}$).

Then, a single bank can settle its own payments at zero cost, while a group of two incurs a positive cost (recall the meaning of threshold K: liquidity is free up to K). That is, for small volumes there are decreasing returns to scale, and hence no incentives to aggregation. See also Figure 1: for low lambdas, expected costs increase slowly i.e. benefits from agglomeration are small.

The outcomes in Figure 4 are clearly *not* what we observe in reality. Encouragingly, using a more realistic matrix (P''') we obtain the more interesting results of Figure 5.

Figure 5: Payments P'''



Type-I costs Type-II costs

Figure 5 shows that, as in the previous example, Type-II costs generate less tiering than Type-I costs. However, given that the banks of matrix P''' differ in their payments, some tiering occurs. Notably, the shape of client-correspondent network is influenced by the shape of the underlying payment network: three correspondents emerge in each of the three 'areas' of more intense exchange.

3.2 Calibrating the model with real data

In this section we calibrate the model using real data, to test the model's ability to yield realistic predictions.

Our first task is to reconstruct the underlying matrix of payments. For this, use the Bank of England 2003 CHAPS traffic survey dataset (also used in Becher et al (2007)). This survey samples five days of the payments executed on the UK large value payment system CHAPS, recording both the ultimate payer and payee banks, and the correspondents used to make these payments. Crucially, the survey also asks correspondent banks to report the percentage of internalized payments. This allows us to determine the volume of payments executed by each correspondent bank which do not cross the CHAPS network. We allocate these payments between banks who use the same correspondent in proportion to their outgoing payments over CHAPS, to obtain the matrix of underlying payment intensities P^*.

We want to run simulations under both Type I and Type II costs, the reason being that these two specifications represent two common regimes of liquidity pricing, on the part of central banks. Type-I applies to a situation where a bank pays proportionally to its liquidity usage (as in the US Fedwire system, where the Fed charges an interest rate on overdrafts). Type-II instead applies to systems where liquidity is given against collateral *and* collateral is (essentially) free up to a certain point. A notable example of this is the UK system CHAPS. There, intraday liquidity can be obtained from the Bank of England at a zero interest rate, in exchange for collateral. But, a certain amount of collateral must be held *anyway* for prudential reasons. Hence, up to the amount of this 'sunk cost', UK banks may obtain liquidity essentially for free.

With Type I costs the model requires no calibration beside P^*.[10] However, under Type II costs we have to choose the threshold K (but only this)[11]. We do so considering two different specifications: under the first, K is constant across banks - we call this specification 'absolute threshold'. Under the second, called 'relative threshold', K varies between banks: for bank i it is given by $\alpha\lambda_i$, with $\alpha > 0$ (of course, the two specifications coincide for $K = 0$). It is difficult to determine an estimate of an *absolute* K: the same threshold could be very small relative to the payment activity of large banks, and very large for small banks. So, the absolute-threshold case is somewhat unrealistic; we use it anyway as a benchmark case. More realistic is the specification $K_i = \alpha\lambda_i$. Indeed in the UK system, the collateral held for prudential reasons is proportional to a bank's potential liquidity outflows. What is a realistic α, then? Comparing data on collateral holdings and payment activity for UK banks (both available to the Bank of England), one ends up with an estimate in the range $0.1 - 0.3$. This is the range of α that we use in the simulations.

We thus run simulations for a range of parameters, observing the resulting equilibrium network. Recall that banks are called to make their decisions in a random order so, even for the same set of parameters, simulations may give different results.

[10] With Type-I costs, expected costs (Eq.2) are linear in c. So, c re-scales all offers proportionally, and is irrelevant for the banks' choices.

[11] With Type-II costs, Eq.2) is no longer linear in c. However, it is in f. So we can normalize f (Eq.4) by c, and obtain that only K/c (i.e. K) matters.

Figure 5: Payments P^*, Type-I costs ($K = 0$)

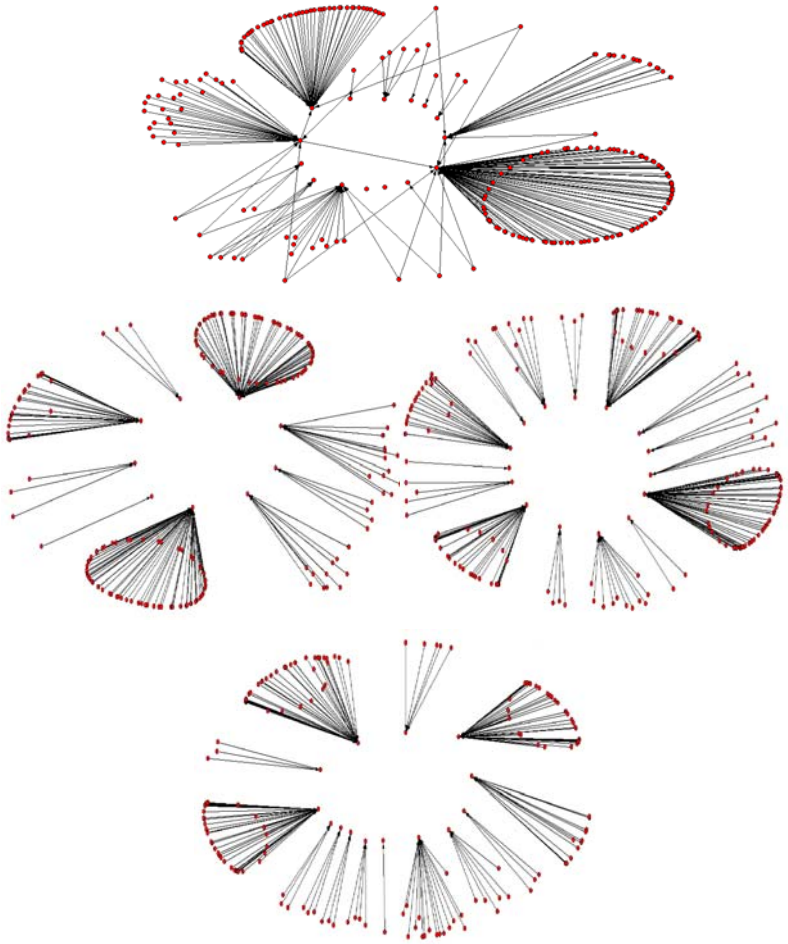


Figure 5 confronts the outcomes of three typical simulations with a map of the real client-correspondent relationships in CHAPS. If it were not that a few real life banks have more than one correspondent, the networks in Figure 5 would appear quite similar. The following analysis shows that the similarity goes beyond an eyeball test.

Tables 1 and 2 show (average) results obtained in the simulations, for different parameter values. The columns report: the number of direct participants obtained in the simulations; the number of real-life CHAPS participants correctly identified; the proportion of internalized payments; and Gini coefficients for the shares of payments, and for the number of customers, accounted for by each correspondent. For comparison, Table 3 reports the same statistics for the real CHAPS.

Table 1: Simulations - Absolute threshold[12]

K	*N. corr.*	*Correct id.*	*Internal. pay.*	*Gini pay.*	*Gini cust.*
0 (Type I)	12.3	5.8	28%	0.58	0.58
5	18.2	6.3	25%	0.70	0.64
10	33.2	6.0	41%	0.80	0.71
20	60.4	9.8	29%	0.76	0.61

Table 2: Simulations - Proportional threshold[13]

$\boldsymbol{\alpha}$	*N. corr.*	*Correct id.*	*Internal. pay.*	*Gini pay.*	*Gini cust.*
0 (Type I)	12.3	5.8	28%	0.58	0.58
0.125	12.4	6.5	48%	0.63	0.67
0.15	12.1	5.9	48%	0.62	0.67
0.175	24.2	6.0	26%	0.75	0.75
0.25	35.4	6.4	17%	0.80	0.72

Table 3: Real system

	N. corr.	*Correct id.*	*Internal. pay.*	*Gini pay.*	*Gini cust.*
CHAPS	14	N/A	33%	0.61	0.69

With an absolute K, the number of direct members of the payment system increases sharply as the threshold rises. By contrast, when the threshold is proportional to the bank's payment volume, increases in α at first do little to encourage direct participation until a critical point is passed (somewhere between thresholds of 15% and 17.5% of payments). This pattern is consistent with the fact that most banks make very small amounts of payments; so, even for small values of an absolute threshold, they are able to settle all payments for free as direct members.

Somewhat surprisingly, the number of real-life CHAPS members identified by the simulations as direct payment system participants is fairly low (typically around 6 of the 14 actual members) and does not increase with rising payment system participation. The banks which are correctly identified as direct participants are, however, the core participants which account for the vast majority of payment flows in CHAPS. Other CHAPS members are only rarely identified as direct payment system members. This suggests that either our model of their liquidity-saving decision is missing something, or that these banks have additional motives for becoming direct CHAPS members (possibilities include historical accident, interdependencies with other payment and securities settlement systems, or a desire to offer sterling correspondent banking services to overseas customers). Alternatively, this could be an artefact of the short period of sample data which we have available to build the matrix P^* - particularly if payments made by some banks show significant seasonality.

[12] Twenty simulations were run for rows 1, 2. Five simulations were run for rows 3, 4.
[13] Twenty simulations were run for rows 1-3. Five were run for rows 4, 5.

Both Gini coefficients rise along with the number of participants. This is due to the small size of the participants drawn into direct membership as liquidity becomes cheaper. With an absolute threshold, the best fit is obtained with $K \simeq 10$, which however generates too many direct participants. A relative threshold specification performs better: for α between 0.15 and 0.175, we obtain a relatively good fit for both the number of participants, and the Gini coefficients.

Under an absolute threshold, there is no clear trend in the percentage of payments which is internalised, as the threshold is raised. Under variable thresholds, the amount of internalisation initially rises as the threshold is increased from zero. Variable thresholds at medium levels give high amounts of internalised payments, but unlike the case of absolute thresholds these drop rapidly as the number of direct participants increases. This difference in behaviour makes sense, as under absolute thresholds: when the threshold rises, smaller banks will be the first to benefit and so opt to participate directly in the payment system. Under variable thresholds larger banks benefit as well.

Summing up: the fixed- and relative-threshold specifications give both a reasonably good fit to the data for low K (the *similarity* is of course expected, as the two models coincide for $K = 0$; the relative *goodness* of fit is instead the pleasing result). Of the two specifications, we prefer the second, as it represents more faithfully liquidity costs in CHAPS. And encouragingly, when K is increased so that the two specification diverge, the relative-K model outperforms the absolute-K model, which indeed produces the counterfactual result that many small banks become direct members. The best fit is attained for an α between 0.15 and 0.175, which is consistent with estimates of real αs. However, judging from the amount of internalised payments (which appears to depend non-monotonically on α), a reasonable fit is also obtained for a very low α (close to Type-1 costs).

4 Conclusions

This paper studies the influence of liquidity costs on the degree of tiering in LVPSs. We formally model the 'netting' and 'liquidity pooling' effects, exploring how these can shape the client-correspondent network of a payment system. The model is extremely parsimonious, requiring essentially two inputs: a matrix of payments (P), and a liquidity cost function (f). Still, when a simple parametrization is performed using data on the main UK payment system, it produces rather complex networks, which bear an encouraging resemblance with what is observed in reality.

Our results suggest that in a regime of free, collateralised intraday credit (such as that operating in the UK), the amount of available collateral has a non-monotonic effect on tiering, first increasing and then decreasing it. This invites the conjecture that a central bank operating such a regime which wished to reduce the level of tiering could do so by broadening the range of collateral which it accepts in exchange for intraday liquidity - especially to include assets

held by smaller banks for reasons unrelated to payments activity, which would therefore provide a cheap source of liquidity. Such a policy would, of course, need to be weighed against its other risks and benefits.

The simulations also suggest that, if liquidity is charged proportionally to its use (as when applying an interest rate), economies of scale set in rapidly, stimulating tiering. If instead the price of liquidity is low for small amounts of liquidity, then higher, such non-convexities weaken the economies of scale and the 'pooling effect', thus making it relatively more convenient to join directly the system. This result may seem at odds with reality: the UK system CHAPS is highly tiered, and yet liquidity is provided against collateral. Instead the Fed charge an interest rate for liquidity, and US system Fedwire is very little tiered. This is probably caused by the fact that, in our model, liquidity needs are generated by 'Poisson payments' - an acceptable modellization for CHAPS, but less good for Fedwire. Indeed, the intensity of CHAPS payments is rather constant throughout the day (see Becher *et al.* (2008)). Instead, Fedwire payments spike at *one* particular point in the day (McAndrews *et al.* (2000), Armantier *et al.* (2008)), suggesting that payments are managed differently in the two systems.

Starting from this observation, further work could look at different specifications for the payments' arrival process. Similarly, one could allow for strategic payment behaviour on the part of banks - which would again alter the time profile of payments and liquidity needs, and so be equivalent to a different assumption on the payment arrival process. In a different line, one could incorporate other forms of costs, such as credit risk, into banks' decisions. Finally, the empirical analysis carried out here could be applied to other LVPSs, perhaps applying formal tests to measure the congruence of the 'artificial' networks with the 'real' ones.

5 Appendix I - derivation of Eq. 2)

Banks in our model make unit payments, at times determined by a Poisson arrival process with parameter λ (which implies that λ is the average number of payments per day). We are interested in the expectation of the maximum of the bank's net debit position, which determines their liquidity costs for the day.

Denote the net debit position of a bank at time t by L_t, and the total number of payments made in the day by N. L_t is a continuous time Markov process. To calculate a bank's expected liquidity cost we need to find

$$E(f(\max_{t\in[0,1]} L_t)) = E(E(f(\max_{t\in[0,1]} L_t|N=n))$$

where the equality follows from the law of iterated expectations.

We can find the inner expectation by considering the jump chain of L_t (see e.g. Norris (1997) for definition), which is simply a symmetric random walk. It is a well-known result, proved using Andre's reflection principle, that

$$\begin{aligned} P(\max_{t\in[0,1]} L_t &= m|N=n) = \\ P(L_1 &= m|N=n) + P(L_1 = m+1|N=n) = \\ &\frac{1}{2^n}\left[\binom{n}{\frac{n+m}{2}} + \binom{n}{\frac{n+m+1}{2}}\right] \end{aligned}$$

Note that L_1 must be odd if N is odd, and even if N is even, and so one of $P(L_1 = m|N=n)$ and $P(L_1 = m+1|N=n)$ is zero. Thus,

$$E(f(\max_{t\in[0,1]} L_t|N=n) = \frac{1}{2^n}\sum_{m=0}^{m=n} f(m)\binom{n}{\lceil\frac{n+m}{2}\rceil} = F(n)$$

(where $\lceil x\rceil$ is the smallest integer larger than x) and

$$E(f(\max_{t\in[0,1]} L_t)) = E(E(f(\max_{t\in[0,1]} L_t|N=n)) = \sum_{n=0}^{\infty}\frac{\lambda^n}{n!}\frac{1}{2^n}\sum_{m=0}^{n} f(m)\binom{n}{\lceil\frac{n+m}{2}\rceil} = C(\lambda)$$

To show that this is increasing, consider

$$C'(\lambda) = e^{-\lambda}\left(\sum_{n=0}^{\infty}\frac{n\lambda^{n-1}}{n!}\frac{1}{2^n}F(n) - \sum_{n=0}^{\infty}\frac{\lambda^n}{n!}\frac{1}{2^n}F(n)\right) = e^{-\lambda}\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{(n-1)!}\frac{1}{2^n}\left[F(n)-F(n-1)\right]$$

So $C(\lambda)$ is increasing if $F(n)$ is. Now,

$$\begin{aligned} F(n)-F(n-1) &= \sum_{m=0}^{m=n} f(m)\binom{n}{\lceil\frac{n+m}{2}\rceil} - \sum_{m=0}^{m=n} f(m)\binom{n-1}{\lceil\frac{n-1+m}{2}\rceil} = \\ &f(n) + \sum_{m=0}^{m=n} f(m)\left[\binom{n}{\lceil\frac{n+m}{2}\rceil} - \binom{n-1}{\lceil\frac{n-1+m}{2}\rceil}\right] \end{aligned}$$

which is positive, since by Pascal's rule $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$.

6 Appendix II - proof of lemma 1

For a given network $\Xi = \left\{S^1, S^2, S^3 ...\right\}$, consider the network's total costs $TC = \Sigma_i C\left(S^i\right)$. We prove that TC is a Lyapunov function for the protocol: every time the network changes, TC decreases. So TC eventually reaches a minimum, at which point the network stops changing.[14] Behind this, the constant application of the NBR to determine both offers and penalties: a new correspondent relationship is established only if some 'extra profit' is realized.

Suppose then k accepts i's offer to join S; there are two cases: a) k is correspondent (for say P), b) k is client of $w \neq k$.

Case a) If k accepts i's offer it must be $\omega > O_i + O_k$ (NBR). That is:

$$\begin{aligned} -C\left(S \cup k\right) + \sum_{r \in S \setminus i} q_{ir} - X_k &> \left(-C(S) + \sum_{r \in S \setminus i} q_{ir}\right) + \left(\sum_{r \in P \setminus k} q_{kr} - C(P)\right) \Rightarrow \\ C\left(S \cup k\right) - \sum_{r \in S \setminus i} q_{ir} + \sum_{r \in P \setminus k} \left[C\left(r\right) - q_{kr}\right] &< C(S) - \sum_{r \in S \setminus i} q_{ir} + C(P) - \sum_{r \in P \setminus k} q_{kr} \Rightarrow \\ C\left(S \cup k\right) + \sum_{r \in P \setminus k} C\left(r\right) &< C(S) + C(P) \end{aligned}$$

On the l.h.s. there are the costs of all banks affected by k's decision[15] when k joins S and P is disbanded. On the r.h.s., the costs that would obtain otherwise, with S and P unchanged. Thus, if k joins S, TC falls. The same inequality can also be obtained from $q_{ik} < -O_k$ (k prefers i's offer to its own profits as correspondent).

Case b) Suppose k is a client of $w \neq k$, correspondent for some P. If k accepts i's offer, this must be more convenient than w's offer:

$$\begin{aligned} d_{ik} &< d_{wk} \Rightarrow \\ \frac{1}{2}\left[-\omega\left(i,k\right) + O_i - O_k\right] &< \frac{1}{2}\left[-\omega\left(w,k\right) + O_w - O_k\right] \Rightarrow \\ -\omega\left(i,k\right) + O_i &< -\omega\left(w,k\right) + O_w \Rightarrow \\ \left(C\left(S \cup k\right) - \sum_{r \in S \setminus i} q_{ir}\right) + \left(-C\left(S\right) + \sum_{r \in S \setminus i} q_{ir}\right) &< \\ \left(C\left(P \cup k\right) - \sum_{r \in P \setminus w} q_{ir}\right) + \left(-C\left(P\right) + \sum_{r \in P \setminus w} q_{ir}\right) &\Rightarrow \\ C\left(S \cup k\right) - C\left(S\right) &< C\left(P \cup k\right) - C\left(P\right) \Rightarrow \\ C\left(S \cup k\right) + C\left(P\right) &< C\left(P \cup k\right) + C\left(S\right) \end{aligned}$$

On the l.h.s. there are the costs when k joins S instead of P; on the r.h.s., the costs that obtain otherwise. Thus again, if k joins S, TC fall. The same

[14] With a finite number of banks and networks, TC takes on a finite number of values.

[15] Bank K's acceptance affects only the costs of i, k, and of the banks in P.

inequality would obtain from $\omega > O_i + O_k$ i.e. for k to join i, the surplus to share must exceed the sum of the outside options.

References

[1] Armantier, O, Arnold J and McAndrews J (2008), "Changes in the Timing Distribution of Fedwire Funds Transfers", forthcoming on the FRBNY Economic Policy Review.

[2] Angelini, P (1998), "An Analysis of Competitive Externalities in Gross Settlement Systems", Journal of Banking and Finance n. 22, pages 1-18.

[3] Bech, M L and Garratt, R (2003), "The intraday liquidity management game", Journal of Economic Theory, Vol. 109(2), pages 198-219.

[4] Becher, Soramaki and Millard (2007), "The topology of CHAPS payments" - forth. Bank of England wp.

[5] Becher C., Galbiati M., Tudela M. (2008), "The Timing and Funding of CHAPS Sterling Payments", forthcoming on the FRBNY Economic Policy Review.

[6] Chapman J, Chiu J, Molico M, "A model of settlement networks", mimeo.

[7] Galbiati M and Soramaki K (2007), "An agent-based model of a payment system", forth. BoE wp.

[8] Harrison, S, Lasaosa, A and Tudela, M (2005), 'Tiering in UK payment systems', Bank of England Financial Stability Review, June, pages 63-72.

[9] Inaoka H., Ninomiya T., Taniguchi K., Shimizu T., Takayasu H. (2004) "Fractal Network derived from banking transactions - An analysis of network structures formed by financial institutions", Bank of Japan Working Paper Series, No. 04-E-04, April 2004.

[10] Jackson, J and Manning, M (2006) 'Central bank intra-day collateral policy and implications for tiering in RTGS payment systems', mimeo.

[11] Kahn, C M and Roberds M (2005) 'Payments settlement: tiering in private and public systems', University of Illinois and Federal Reserve Bank of Atlanta, mimeo. http://www.bankoengland.co.uk/financialstability/futureofpayments/kahnroberdsBOE.pdf

[12] Lai A, Chande N, O'Connor S, "Credit in a tiered payment system", Bank of Canada working paper, 2006-36.

[13] Lasaosa A, Tudela M, "Risks and efficiency gains of a tiered structure in large-value payments: a simulation approach", BoE forthcoming WP.

[14] Lublóy Á (2006), "Topology of the Hungarian large-value transfer system", Magyar Nemzeti Bank, Occasional Papers 57, 06.

[15] McAndrews, J. and Rajan, S. (2000), 'The Timing and Funding of Fedwire Funds Transfers', Federal Resever Bank of New York Economic Policy Review 6 (2), 17-32.

[16] Norris J R. (1997), "Markov Chains", Cambridge Series in Statistical and Probabilistic Mathematics, CUP.

[17] Soramäki, Kimmo, M.L. Bech, J. Arnold, R.J. Glass and W.E. Beyeler (2008). "The Topology of Interbank Payment Flows". Physica A. Vol. 340, pp. 380-394.

[18] Harrison S, Lasaosa A, and Tudela M (2005), "Tiering in UK Payment Systems: Credit Risk Implications". Bank of England Financial Stability Review Issue 19 pp. 63-70.